Using parallel corpora in teaching & research:
The Swedish-Hindi-English & Swedish-Turkish-English parallel corpora

Anju Saxena
Beáta Megyesi
Éva Csató Johanson
Bengt Dahlqvist

Department of Linguistics and Philology
Uppsala University

1. Introduction

Proposed methods for the automatic acquisition of linguistic knowledge by computer potentially allow for the rapid creation of language technology resources with minimal human work, which if realized would be of great help in the case of many less-commonly taught languages. In this connection, it is important to note that, arguably, language technology – and consequently also the application of machine learning methods in language technology – has been shaped by the typological and other traits of the most explored languages, especially English, which is in many respects an atypical language from a linguistic point of view, and quite unlike many other languages. There is, thus, a need to test and refine these methods on a number of structurally different languages, making languages such as Turkish, Swedish and Hindi a good testing ground, allowing us to gain a better understanding of the generality or language-specificness of these methods.

"Supporting Research Environment for Minor Languages" is a research program at the Department of Linguistics and Philology, Uppsala University – financed by the Swedish Research Council and Uppsala University – the aim of which is to provide a research environment for less explored languages by developing parallel treebanks. We are currently working on building a Turkish-Swedish-English and a Hindi-Swedish-English parallel treebanks. The aim of this article is to describe briefly our work with these two parallel treebanks and also to show how these treebanks are a valuable resource in teaching and research. Section 2 describes briefly some important aspects of the two parallel corpora. In section 3 we will describe how corpora can be used in linguistic research and in teaching in general and our experience of using the Swedish-Turkish treebank in teaching in particular.

2. On the Swedish-Hindi-English and Swedish-Turkish-English parallel treebanks

A *parallel corpus* is a bi- or multilingual text material containing original texts in one language and their translations into another language or other languages (or, alternatively, parallel translations from an original in a language not in the corpus; this is normally the case with parallel corpora of Bible texts). Often, parallel corpora are *aligned*, meaning that corresponding units (sentences, phrases, even words) from the different language versions are explicitly linked together. Syntactically annotated corpora are called treebanks.

The focus in language technology has been on English and major Western languages. English has been one of the languages included in many of the existing parallel treebanks. For example,

- The Prague Czech-English Dependency Treebank (Hajic, 2001)
- ISJ-ELAN Slovene-English Parallel Treebank (Erjavec, 2002)

- Swedish-English Parallel Treebank (Ahrenberg, 2007)

Relatively little work has been done on less-commonly taught languages. This is one major reason for including Hindi, Swedish and Turkish in this project. Up until now there has been no parallel corpora involving Swedish-Hindi or Swedish-Turkish.

The Swedish-Hindi-English parallel treebank
Hindi provides a good testing ground for language technology tools. There are, at present, two major Hindi treebanks generally available: the EMILLE treebank (<http://www.emille.lancs.ac.uk/>; Hardie et al. 2006) and the Hindi treebank at IIT Bombay (<http://www.cfilt.iitb.ac.in/>). The IIT Bombay Hindi treebank is a monolingual treebank, consisting of excerpts of texts each around 2 000 words (which can be an issue for discourse related work), and it does not seem to have any linguistic annotation. The EMILLE treebank has both a large monolingual and a smaller parallel treebank part (Hindi-English), but this treebank, too, is not linguistically annotated. Keeping in view our goal of creating a trilingual parallel treebank which is POS-tagged and dependency parsed as well as aligned at word and sentence levels, and, keeping in view our ambition to use this treebank in research in linguistics and in teaching makes these existing Hindi treebanks less useful for our purposes. Therefore decision was made to start from scratch. The Hindi materials which, at present, are included in our treebank are.

- Bible texts: the 4 Gospels
- Texts from the parallel treebank section of the EMILLE project
- The UN Declaration of Human Rights
- A Hindi novel
- Some texts providing information about Sweden (see below)

Hindi texts were typed in manually for lack of usable electronic versions. All texts have undergone semi-automatic cleaning and converting in order to make them conform to the requirements of the annotation tools used. The texts are now available in XML with Unicode (UTF-8) character representation. Initial work has also been done regarding POS tagging, morphological analysis and chunking. MaltParser (Nivre et al., 2006a) has been trained on a syntactically annotated Hindi treebank (Saxena et al 2008).

The following table summarizes the details of the texts included in the Swedish-Turkish-English and the Swedish-Hindi-English parallel corpora. It shows that while some texts are unique to one of the two corpora, a large number of the texts are common in the two treebanks:

TOKENS				
Text	**Swedish**	**English**	**Turkish**	**Hindi**
Pregnancy	1 382	439	1076	1 221
Movement	711	834	616	685
Psychology	348	383	385	330
Retirement	-	-	3 770	4 267
Dublin	496	564	451	469
UN Declaration of Human Rights	1 911	2 106	1 831	1 604
What is unicode	514	626	539	424

Gospel of Luke	32 238	30 621	-	32 238
Gospel of Matthew	19 564	29 274	-	29 247
Gospel of Mark	18 872	18 481	-	18 888
Gospel av John	24 209	24 907	-	24 625
Total	**133 568**	**108 235**	**288 701**	**162 302**

Important considerations while building these parallel corpora have been that:

- The texts are high quality texts, where the translations are done by trained translators. Compiling a proper text treebank entails a much greater amount of work than merely collecting any kind of text that you can lay your hands on, especially where other text types than newstext are difficult or impossible to acquire in electronic form. Texts are morphologically and syntactically analyzed

- They are (semi-)automatically aligned at the sentence and word levels

- Special attention has also been made to explore and to the extent possible, to use open-source resources and to use the same tools in working with the Swedish-Turkish-English and the Swedish-Hindi-English parallel corpora. This has several advantages. For example, this gives us a chance to examine the usefulness of a given tool across languages. In this way, it provides us a set of tools which are not language-dependent. This, at times, has meant using an existing tool as it is (and we have in fact often been able to do just that), while at other times it has meant, developing an existing tool further.

Since both these parallel corpora have, to a large extent, similar format and they have used similar sets of tools, we will below describe the building of the two corpora by describing in more detail the building of the Swedish-Turkish-English parallel corpus.

The Swedish-Turkish-English parallel treebank
The Swedish-Turkish parallel treebank, which is currently under development and which has been previously described (Megyesi et al. 2006; Megyesi & Dahlqvist 2007; and Megyesi et al. 2008) contains syntactically annotated parallel texts with various annotation layers from part-of-speech tags and morphological features to dependency annotation where each layer is automatically annotated, the sentences and words are aligned, and partly manually corrected.

The treebank material is processed automatically by using various tools making the annotation, alignment and manual correction easy and straightforward for users with less computer skills. This is necessary, as our ambition is to allow researchers and students of particular languages to enlarge the treebank by automatically processing and correcting the new data by themselves.

In order to build the treebank automatically, we use a basic language resource kit (BLARK) for the particular languages and appropriate tools for the automatic alignment and correction of data.

First, the original materials received from the publishers in various formats are cleaned up. For example, rtf, doc, and pdf documents are converted to plain text files. After cleaning up the original data, the texts are processed automatically by using tools for formatting, linguistic annotation and sentence and word alignment.

During formatting, the texts are encoded using UTF-8 (Unicode) and marked up structurally using XML Treebank Encoding Standard (XCES). The text files are processed by various tools in the BLARKs developed for each language separately. A tokenizer is used to split the text into tokens such as words and punctuation marks. Sentence segmentation is also performed to break the texts into sentences.

Once the sentences and tokens are identified, the data is linguistically analyzed. We use several annotation layers for the linguistic analysis, first on a morphological level, then on a syntactic level. The annotation and the labels for the linguistic analysis are *de facto* standard for the involved languages. For the linguistic annotation, external morphological analyzers, part-of-speech taggers and syntactic dependency parsers are used which are trained on annotated treebanks developed for the specific languages. For example, for Swedish we use the Stockholm Umeå Treebank tag set (SUC 1997) for the morpho-syntactic annotation and the functional annotation of Talbanken05 (Nivre et al. 2006b), while we derive the linguistic annotation from the Metu-Sabancı Turkish Treebank (Oflazer et al. 2003) for the syntactic analysis of Turkish. For English, we use the Penn Treebank tag set.

The Swedish and English texts are morphologically annotated with the open source HunPoS tagger (Halacsy, et al. 2007). The tokens are annotated with parts of speech and morphological features and are disambiguated. The results for the morphological annotation of Swedish show an accuracy of 96.6% (Megyesi 2008). The Turkish material is morphologically analyzed and disambiguated using a Turkish analyzer (Oflazer 1994) and a disambiguator (Yuret & Türe 2006) with an accuracy of 96%. The English data contains less error, approximately only 2%-3%.

The other linguistic layer contains information about the syntactic analysis. We use dependency rather than constituent structures, as the former has been shown to be well suited for both morphologically rich and free word order languages such as Turkish, and for morphologically simpler languages, like Swedish. The English, Swedish and the Turkish data are annotated syntactically using MaltParser (Nivre et al. 2006a), trained on Penn Treebank and Talbanken05 (Nivre et al. 2006b) and on the Metu-Sabancı Turkish Treebank (Oflazer et al. 2003). The annotation includes approximately 15%-20% errors, depending of the language, which need to be manually corrected.

The sentences and words in the languages are aligned automatically by using standard techniques, such as the length-based approach (Gale and Church 1993) for sentence alignment, and the clue alignment approach (Tiedemann 2003) with the toolbox for statistical machine translation GIZA++ (Och and Ney 2003) for word alignment. The aligned sentences are manually corrected by a student who speaks both languages. We automatically compare the links before and after the manual correction and the user gets statistics about the differences. The results show that between 67% and 94% of the sentences are correctly aligned by the automatic sentence aligner depending on the text type. We are currently working the automatic correction of word alignment in the syntactic trees.

In addition, we visualize the treebank in different ways without showing the structural markup when used, for example, in teaching.

3.1 Use of treebanks in research and in teaching

There has been a growing interest in using natural language corpora in teaching and in research, partly due to the growing availability of computer-readable linguistic corpora, and partly due to an increase in examining language in its natural context as opposed to investigating constructed language examples in isolation. Researchers, teachers and students now have access to different types of language corpora to discover facts about language; for example, which words are the more frequently used words in a language or a language-type, in which context they predominantly occur and which grammatical patterns are associated with a particular linguistic item (Ghadessy *et al* 2000). There have been two primary approaches for the use of treebanks in language teaching/learning: the "COBUILD approach" and the Data-Driven Learning approach.

Until recently, the COBUILD approach was the predominant approach. Corpora, in this approach, are used by researchers and producers in building dictionaries and other language learning materials. Traditionally it has been very large corpora which have been used for this purpose. Further, within this approach, the user (a student, for example) receives results of a project involving corpora as end-products (for example, in the form of a language learning packet). Learners do not get to use the corpora themselves in order to come up with their own analyses and learn from that. Another limitation of this approach has been its limited access. Up until quite recently access to the results of such works has been limited, primarily because of the high cost of such language-learning tools.

In the Data-Driven Learning approach students use corpora directly in their own learning. They use the corpora, for example, to discover linguistic patterns and to organize linguistic patterns which they observe, arriving at generalizations inductively and verifying deductive rules. Such exposure to corpora provides students the chance not only to extract relevant examples of one or the other linguistic structures, but also provides them material for discussion when they find gaps, to verify and extend their hypothesis and to arrive at generalizations. In favour of the Data-Driven Learning approach, Tim Johns (1991) states:

> What distinguishes the data-driven learning approach is the attempt to cut out the middleman … and give direct access to the data so that the learner can take part in building up his or her own profiles of meanings and uses. (Johns 1991:30 in Aston 1997)

Johns (1991) mentions three phases in the Data-Driven Learning:

- observation
- classification
- generalization

One advantage of using corpora in teaching is that instead of learning about linguistic theories in vacuum (*a more passive learning method*, where facts are fed to students in form of lectures), students have a chance to test these theories themselves against these corpora and learn about these theories or concepts for themselves (*a more active learning method*). When treebanks are used by students as part of their learning, distinction between teaching and research is "blurred", as students, by discovery proceedure (thus, research), learn things for themselves (Knowles 1990). The use of corpora in teaching can, in this way, affect both teachers' as well as students' role. This approach is as equally relevant in a classroom set-up as in self-study situations.

The gap between the COBUILD and DDL approaches is, however, getting smaller. More access to corpora (especially for non-commercial purposes) provides better (pre-)conditions to use them in producing language learning tools as well as in using them directly in teaching/learning.

3.2 Using Swedish-Turkish-English treebank in a teaching enviroment
The aim of the Swedish-Turkish treebank (STPC) is to provide Swedish speaking students and researchers with easily accessible annotated linguistic data on Turkish. The corpus is now being completed with English texts. The webbased STPC can be used both by regular and distance students in their data-driven acquisition of new vocabulary items and their usage. It functions also as a learning platform for and for testing hypotheses concerning the morphological and syntactic aspects of Turkish grammar. Further, it helps the students to practice translation between Swedish and Turkish. All this is possible due to the fact that the Swedish-Turkish parallel texts are available in annotated form. The annotations, on request, are visualized in pop-up windows. The morphological analyses are given at present in clumsy, parser-generated formulas but will in the near future be substituted by labels in more intelligible forms based on the grammatical terms employed in textbooks and in the Turkish Suffix Dictionary (Csató and Nathan 2003). The interface for displaying syntactic information is not ready yet.

A search tool assists the students to create concordance lists. They can search for whole words, beginnings of words, parts of words or ends of words in Turkish or Swedish. The concordance lists display whole sentences in which the target item appears and it is highlighted. The selected sentences are aligned with their translational equivalents. This form of displaying the linguistic data is much more suitable for learning than KWIC lists in which only the immediate environment of the target item is shown.

Search

Sökresultat

Text: Vita Borgen
Söksträng: artık
Antal funna meningar: 88 |
Antal förekomster: 89

num	swedish	turkish
28	Jag följde dem en tid , men blev uttröttad , det kom svar från italienska universitetet som gjorde slut på mitt hopp . Även efterforskningarna jag gjorde på kyrkogårdarna i Gebze , Cennethisar och Üsküdar där jag letade efter författarens namn blev ofruktsamma . Jag slutade jaga spår och tog med författaren i encyklopedin med hjälp av uppgifterna i boken .	Bir süre onların peşinden gittim , ama bıkmıştım artık , mektup yağmuruna tuttuğum İtalyan üniversitelerinden umut kırıcı cevaplar geliyordu : Gebze , Cennethisar ve Üsküdar mezarlıklarında yazarın kitabın kendisinden çıkan , ama üzerinde yazmayan adına dayanarak yaptığım araştırmalar da başarısız çıkmıştı : İz sürmeyi bıraktım , ansiklopedi maddesini hikâyesinin kendisine dayanarak yazdım .
55	Vår kapten började hoppas när han såg hur de två andra skeppen slingrade sig fram mellan de turkiska fartygen och försvann i dimman , och till slut fick han , efter våra påtryckningar , mod att låta piska slavarna , men nu var det för sent ; dessutom kunde inte ens piskorna ta de av frihetslängtan upphetsade slavarna att lyda .	Öteki iki geminin arasından sıyrılıp sisin içinde kaybolduğunu görünce kaptanımız umutlandı , bizim de zorumuzla esirleri sıkıştırmaya cesaret edebildi , ama geç kalmıştık artık ; üstelik özgürlük tutkusuyla heyecanlanan kölelere kırbaçlar da söz geçiremiyordu .
105	Folk hade hört att jag var läkare , jag behandlade inte bara slavarna som ruttnade i vårt fängelse utan även andra .	Yalnız zindanda çürüyen kölelere değil , hekim olduğumu işiten başkalarına da bakıyordum artık .
141	Jag fick fortsätta arbeta men nu behandlades jag förmånligt av slavdrivarna .	Gene işe çıkarılıyordum , ama esirbaşları artık kayırıyorlardı beni

Such lists are used to find frequent patterns of usage, transformational equivalents, different meanings of polysemic words, translational equivalents of Turkish grammatical categories, etc. Different types of exercises are designed and published in the Internet. Students in Turkic languages also use STPC while writing their theses. Bergdahl (2006) studied the meanings of the Turkish word *gölge* 'shadow' and the corresponding Swedish word *skugga*. Dadasheva (2005) investigated how the Turkish indirective category marked by *-mIş* / *imiş* is translated into Swedish and Russian. Hedman and Uyghur (2009) compared the meanings of the

Swedish and Turkish verbs 'give', 'do' and 'make'. Haktanır (2006) reviewed the ambiguous Turkish morphological forms in one of the parallel texts and described different types of morphological ambiguities.

Apart from using STPC in learning environments, it is also being used by researchers. One example of which is the article *Rendering evidential meanings in Turkish and Swedish* (Csató 2009), which examined the Turkish evidential category of indirectivity and the less grammaticalized or lexical strategies in Swedish to render evidential nuances. The description of the strategies used in the two languages was complemented with an analysis of data in one of the parallel Turkish-Swedish texts. It was found that although Swedish has means to express evidential nuances these were much less used in the Swedish translations than expected. The article describes several reasons for this. One might be that the Turkish category allows three different types of reading. This ambiguity is significant in certain texts. The Swedish devices may render a particular evidential nuance but not the whole range of ambiguity of the Turkish forms.

In short, the parallel corpora in general and the Swedish-Turkish-English and the Swedish-Hindi-English treebanks, in this way, can be used in a variety of ways in teaching and in research.

References

Ahrenberg, L. 2007. LinES: An English-Swedish parallel treebank. In *Proceedings of Nordiska Datalingvistdagarna*, Nodalida 2007. Tartu, Estonia

Bergdahl, Eva Annika 2006. Shadow. From a relaxing spot to darkness and death**.** A semantic study of how the word shadow is used in Swedish and Turkish. C-uppsats. Batı Dilleri ve Edebiyatı Bölümü, Bogaziçi University & Department of Linguistics and Philology, Uppsala University

Csató, Éva Á. 2009. Rendering evidential meanings in Turkish and Swedish. In: Éva Á. Csató et al (eds.) *Turcological letters to Bernt Brendemoen*, 77-86. Oslo: Novus

Csató, Éva Á. & Nathan, David. 2003. Turkish suffix dictionary. http://www.dnathan.com/language/turkish/tsd/

Dadasheva, Sabina 2005. *Den turkiska indirektiva kategorin. En undersökning av återgivningen av den turkiska indirektiva kategorin i ryska och svenska autentiska översättningar*. C-uppsats. (In Swedish.) Department of Linguistics and Philology, Uppsala University

Gale, W. A., and K. W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics,* 19.1: 75-102

Ghadessy, Mohsen, Alex Henry and Robert L. Roseberry. 2000. Introduction. In: Mohsen Ghadessy, Alex Henry and Robert L. Roseberry (eds.) *Small corpus studies and ELT. Theory and practice*, xvii-xxiii. Amsterdam/Philadelphia: John Benjamins

Haktanır, Murat 2006. *Orhan Pamuk'un Beyaz Kale adlı eserinde çok anlamlılık*. C-uppsats. (In Turkish.) Department of Linguistics and Philology, Uppsala University

Hedman, Merih 2009. *Verbet 'göra' i svenska och turkiska. C-uppsats*. (In Swedish.) Department of Linguistics and Philology, Uppsala University

Hajič, J., Hajičova, E., Pajas, P., Panevova, J., Sgall, P., and Vidova-Hladka, B. 2001. *Prague dependency treebank 1.0* (final production label). CDROM CAT: LDC2001T10., ISBN 1-58563-212-0.

Halácsy, P., A. Kornai, and Cs. Oravecz. 2007. Hunpos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*,

Companion volume, proceedings of the demo and poster sessions, 209–212, Prague, Czech Republic.

Hardie, Andrew, Paul Baker, Tony McEnery & B. D. Jayaram. 2006. Corpus-building for South Asian languages. In Anju Saxena and Lars Borin (eds.) *Lesser-known languages of South Asia. Status and policies, case studies and applications of Information Technology*, 211-242. Berlin: Mouton de Gruyter

Aston, Guy. 1997. Small and large corpora in language learning. http://www.sslmit.unibo.it/guy/wudj1.htm

Johns, T. 1991. Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (eds.), *Classroom concordancing. English language research journal* 4, 1-16

Knowles, Gerald. 1990. The use of spoken and written corpora in the teaching of language and linguistics. *Literary and linguistic computing. Journal of the Association for literary and linguistic computing*. 5.1: 45-48

Megyesi, B. 2008. The open source tagger HunPoS for Swedish. Department of linguistics and philology, Uppsala University

Megyesi, B. B., A. Sågvall Hein, and E. Csato Johanson. 2006. Building a Swedish-Turkish parallel corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC'06). Genoa, Italy

Megyesi, B. B. and B. Dahlqvist. 2007. A Turkish-Swedish parallel corpus and tools for its creation. In *Proceeding of Nordiska Datalingvistdagarna*, NoDaLiDa 2007

Megyesi, B., B. Dahlqvist, E. Pettersson & J. Nivre. 2008. Swedish Turkish Parallel Treebank. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC'08), 470-473. Marocco

Nivre, J., J. Hall & J. Nilsson. 2006a. MaltParser: A Data-Driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC), 2216–2219

Nivre, J., J. Hall & J. Nilsson. 2006b. Talbanken05: A Swedish Treebank with phrase structure and dependency annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC), 1392–1395

Saxena, A., Swaroop Madhyasta, P., and Nivre, J. 2008. Building the Uppsala Hindi Corpus. Poster presentation. SLTC 2008. Stockholm. <http://www.speech.kth.se/sltc2008/>

Oflazer, K. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9:2

Oflazer, K., B. Say & Hakkani-Tür. 2003. Building a Turkish treebank. In Anne Abeillé (ed.) *Treebanks: Building and using parsed corpora*, 261–277. Kluwer

Och, F. J. & H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.1: 19-51

SUC. Department of Linguistics, Umeå University and Stockholm University. 1997. SUC 1.0 Stockholm Umeå Corpus, Version 1.0.

Tiedemann, J. 2003. *Recycling translations – Extraction of lexical data from parallel corpora and their applications in Natural Language Processing*. PhD Thesis. Uppsala University

Uyghur, D. 2009. *Semantiken av turkiska verbet ver- 'ge' och dess motsvarighet i svenska*. (In Swedish.) Department of Linguistics and Philology, Uppsala University

Yuret, D. & F. Türe. 2006. Learning morphological disambiguation rules for Turkish. In *Proceedings of HLT NAACL'06*, 328-334. New York